Multidimensional Interval-Data: Metrics and Factorial Analysis

Francesco Palumbo[1]* and Antonio Irpino[2]

[1] Department of Economics and Finance - Università di Macerata
Via Crescimbeni, 20
I-62100 Macerata, Italy
(e-mail: francesco.palumbo@unimc.it)

[2] Department of Business Strategies and Quantitative Methods - Seconda Università di Napoli
Corso Gran Priorato di Malta
I-80126 Capua, Italy
(e-mail: irpino@unina.it)

Abstract. Statistical units described by interval-valued variables represent a special case of Symbolic Objects, where all descriptors are quantitative variables. In this context, the paper presents two different metrics in $\mathbb{R}^p$ for interval-valued data that are based on the definition of the Hausdorff distance in $\mathbb{R}$. Hausdorff distance in $\mathbb{R}^p$ (for any $p \geq 1$) is a L_∞ norm between pairs of closed sets. However, when $p > 1$ the problem complexity leads towards the definition of L_2 norms approximating as well as possible the Hausdorff distance. Given a set of n units described by p interval-valued variables, we compute and represent the distances over factorial planes that are defined by factorial analyses that are consistent with the two distance measure definitions.

Keywords: Factorial Analysis, Hausdorff distance, Interval Data.

1 Introduction

Let $\mathbf{\Omega} = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be a set of individuals with description in the space $\mathbb{IR}^p$, where $\mathbb{IR}^p$ indicates the p-dimensional space of the closed subsets in $\mathbb{R}$. The individuals can be modeled as Symbolic Objects (SO) described by interval descriptors. Interval data represent a special case of set-valued data, where the sets are compact and identified by ordered couples of values: $[a] = [\underline{a}, \overline{a}] \subset \mathbb{R}$, which correspond to the interval bound values [Hickey *et al.*, 2001]. The generic $n \times p$ interval data matrix $[\mathbf{A}]$ has general term $[a]_{i,j}$, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$ indicate the generic statistical unit and the generic descriptor, respectively. The general term $[a]_{i,j}$ can also be represented as the *midpoint* $a^c_{i,j}$ and *range* (or *radius*) $a^r_{i,j}$ notation: $[a]_{i,j} =$

* *Acknownledgements:* The paper was financially supported by PRIN2003 grant: "*Multivariate Statistical and Visualization Methods to Analyze, to Summarize, and to Evaluate Performance Indicators*" research unit of University of Macerata, responsible F. Palumbo.

$[\underline{a}_{i,j}, \overline{a}_{i,j}] = [a^c_{i,j} - a^r_{i,j}, a^c_{i,j} + a^r_{i,j}]$. Midpoints and ranges are respectively defined by:

$$a^c = \frac{1}{2}(\overline{a} + \underline{a}), \qquad a^r = \frac{1}{2}(\overline{a} - \underline{a}).$$

In the midpoints/ranges notation, the matrix $[\mathbf{A}]$ is split in the matrices $\mathbf{A}^c$ and $\mathbf{A}^r$ that are called center and range matrix, respectively.

The interval (data) arithmetic has a wide specialized literature, see [Alefeld and Mayer, 2000] for an exhaustive survey. However, the direct treatment of interval-variables in statistics is limited to very few cases, this occurs because the computation of the variance-bounds is an *NP*-hard problem and does not have approximate solutions [Xiang *et al.*, 2004]. It is worth noticing that this aspect involves also the Principal Component Analysis (PCA) and factorial analysis, more generally.

Facing the problem from a geometric point of view and starting from different definitions of distance between intervals, many authors have proposed different approaches to the factorial analysis for interval data (see [Cazes *et al.*, 1997], [Lauro and Palumbo, 2000], [Lauro and Palumbo, 2005], [Giordani and Kiers, 2004]). Generally, a distance between intervals takes into account only some representative points. Cazes et *al.* and Giordani and Kiers based their analysis on the distance between the interval bounds (*vertices*); Lauro and Palumbo proposed a distance measure based on the interval centers and *radii* (or ranges). However, there exist many distance definitions for interval data and more generally for set-valued data; given any general function of distance or proximity, it is possible to arrange a $n \times n$ matrix on which to perform a MultiDimensional Scaling (MDS) analysis and to represent the SO as points in the reduced space.

Dealing with punctual data, a statistical unit is represented by a dimensionless point in $\mathbb{R}^p$ $\forall p$; whereas, the geometric nature of a closed subset ω_i in $\mathbb{R}^p$ varies according to p; it is a segment if $p = 1$, a parallelogram if $p = 2$ a parallelepiped when $p = 3$ and, more generally, a parallelotope when $p > 3$, where $\omega_i = ([a]_{i,1}, [a]_{i,2}, \ldots, [a]_{i,p})$ indicates the generic subsets in $\mathbb{R}^p$, $\forall p \geq 1$.

Differently from the MDS, our aim is to represent the distances but also the *size* and *shape* of the SO [Lauro and Palumbo, 2005].

In section 2 we shall introduce the Hausdorff metric and in section 3 we shall present two distances for interval valued data in $\mathbb{IR}^p$, both of them are derived from the Hausdorff notion of distance. Section 4 presents an application of the two distances on the *Italian peppers* data set. Distances and SO sizes and shapes are represented over factorial planes by means of two factorial analyses; section 5 closes the paper.

2 Distance measures in $\mathbb{IR}^p$

In this section we present the Hausdorff metric for interval data and we introduce two different generalizations in the $\mathbb{IR}^p$ space. We shall show that these distances are good approximations of the Hausdorff distance in $\mathbb{R}^p$ and can be easily decomposed in suitable factorial models.

The Hausdorff metric was proposed by Felix Hausdorff in the early of 20^{th} century as a measure of distance between compact subsets in $\mathbb{R}^p$.

Given a metric $d(\cdot)$, the distance *from* a generic point $x \in \mathbb{R}^p$ *to* a closed subset $A \subset \mathbb{R}^p$ is defined as:

$$d(x, A) = \min_{\tilde{a} \in A} d(x, \tilde{a}).$$

Let $\mathcal{H}(X)$ be the space of all non-empty compact subsets of X, the Hausdorff metric on $\mathcal{H}(X)$ is defined on the basis of the following quantities:

$$h(A, B) = \max_{\tilde{a} \in A} d(\tilde{a}, B), \qquad h(B, A) = \max_{\tilde{b} \in B} d(\tilde{b}, A),$$

where $\{A, B\} \in \mathcal{H}(X)$ and $\{\tilde{a} \in A, \tilde{b} \in B\}$.

The Hausdorff distance $H(A, B)$ is defined by:

$$\begin{aligned} H(A, B) &= \max\{\max\{d(\tilde{a}, B)\}, \max\{d(\tilde{b}, A)\} = \\ &= \max\left(h(A, B), h(B, A)\right). \end{aligned} \tag{1}$$

In the special case of $\mathbb{R}$, the Hausdorff distance between two generic intervals is given by: $H(A, B) = \max\{\mid \overline{a} - \overline{b} \mid, \mid \underline{a} - \underline{b} \mid\} = \mid a^c - b^c \mid + \mid a^r - b^r \mid$. It is easy to show that $H(A, B) \geq 0$ and $H(A, B) = H(B, A)$. Moreover, let C be a generic compact subset in $\mathbb{R}$, the triangular inequality $H(A, C) \leq H(A, B) + H(B, C)$ can be easily proved taking into account the definition of distance in (1) [Neumaier, 1990].

3 Generalization of $H(A, B)$ in $\mathbb{R}^p$

The generalization of the Hausdorff distance in $\mathbb{IR}^p$ tends to be very complex as p tends to be large. Readers interested in the properties of the Hausdorff metric in $\mathbb{R}^p$ space may refer to [Braun *et al.*, 2003]. However, when the compact subsets in $\mathbb{IR}^p$ are restricted to some special cases, the Hausdorff metric can be easily generalized. This paper will focus the attention on two special cases: *boxes* and hyperspheres.

3.1 Distance between *boxes*

In order to have a distance measure easy to handle in $\mathbb{IR}^p$, we introduce a measure of distance that generalizes the Minkowski metric.

In the $p-$dimensional space $\mathbb{R}^p$, $\mathcal{H}(X_1, X_2, \ldots, X_p)$ indicates the set of all possible bounded *boxes* (or *parallelotopes*) in the space $\mathbb{IR}^p$.

Given two *boxes* $\{A, B\} \in \mathcal{H}(X_1, X_2, \ldots, X_p)$, the quantity:

$$H(A, B) = \left\{\textstyle\sum_{j=1}^{p} \mid H(A_j, B_j) \mid^{\alpha}\right\}^{\frac{1}{\alpha}} \geq 0, \tag{2}$$

for any $\alpha \geq 1$, is a metric. It is obvious that $H(A, A) = 0 \Leftrightarrow A = A$, $\forall A \in \mathcal{H}(X_1, X_2, \ldots, X_p)$, being $H(A_j, A_j) = 0, \forall j = 1, \ldots, p$. The two following properties of (2) can be easily demonstrated:

i) $H(A, B) = H(B, A)$ (*Symmetry*):
For any $(A, B) \in \mathcal{H}(X_1, X_2, \ldots, X_p)$ is:

$$\begin{aligned} H(A,B) &= \left\{\textstyle\sum_{j=1}^{p} [H(A_j, B_j)]^{\alpha}\right\}^{\frac{1}{\alpha}} = \\ &= \left\{\textstyle\sum_{j=1}^{p} [H(B_j, A_j)]^{\alpha}\right\}^{\frac{1}{\alpha}} = H(B, A) \end{aligned} \tag{3}$$

given the symmetry of $H(A_j, B_j)$ for any $j = 1, \ldots, p$.
For any $A \in \mathcal{H}(X_1, X_2, \ldots, X_p)$ is:

$$H(A,A) = \left\{\textstyle\sum_{j=1}^{p} [H(A_j, A_j)]^{\alpha}\right\}^{\frac{1}{\alpha}} = 0 \tag{4}$$

ii) $H(A, B) + H(A, C) \geq H(B, C)$ (*Triangular inequality*): For any $(A, B, C) \in \mathcal{H}(X_1, X_2, \ldots, X_p)$ under the hypothesis that the distance $H(A_j, B_j)$ satisfies the triangular inequality for any $j = 1, \ldots, p$, this follows from equation (1) (see [Neumaier, 1990] for a complete specification of the metric properties in the $\mathbb{IR}$ space). The following proves the inequality $H(A, B) + H(A, C) \geq H(B, C)$:

$$\begin{aligned} H(A,B) + H(A,C) &= \left\{\textstyle\sum_{j=1}^{p} [H(A_j, B_j)]^{\alpha}\right\}^{\frac{1}{\alpha}} + \left\{\textstyle\sum_{j=1}^{p} [H(A_j, C_j)]^{\alpha}\right\}^{\frac{1}{\alpha}} \geq \\ &\geq \left\{\textstyle\sum_{j=1}^{p} [H(A_j, B_j) + H(A_j, C_j)]^{\alpha}\right\}^{\frac{1}{\alpha}} \geq \\ &\geq \left\{\textstyle\sum_{j=1}^{p} [H(B_j, C_j)]^{\alpha}\right\}^{\frac{1}{\alpha}} = H(B, C), \end{aligned} \tag{5}$$

being $H(A_j, B_j) + H(A_j, C_j) \geq H(B_j, C_j)$ satisfied for any j, according to the Hausdorff metric definition in $\mathbb{R}$.

The distance in $\mathbb{IR}^p$ introduced in (2), for $\alpha = 2$ can also be expressed in terms of centers and *radii*:

$$H(A,B) = \sqrt{\sum_{j=1}^{p} \left[(a_j^c - b_j^c)^2 + (a_j^r - b_j^r)^2 + 2 \mid a_j^c - b_j^c \mid \mid a_j^r - b_j^r \mid\right]}. \tag{6}$$

This notation will be useful when we shall present the factorial model.

3.2 Hausdorff distance between two spheres in $\mathbb{R}^p$

Another distance in $\mathbb{R}^p$, which derives from the Hausdorff metric in $\mathbb{R}$, is given by the distance defined between the spheres inscribing the *parallelotopes*. This distances coincides with the Hausdorff metric when the SO are hypercubes (equal edges).

Before illustrating the distance we prove the following theorem that defines the Hausdorff distance between spheres in the $\mathbb{R}^p$ space.

In this section capital letters $\{A, B, \ldots\}$ indicate spheres in the $\mathbb{R}^p$ space; the general sphere A has center in $A^c = [a_j^c]$ $(j = 1, \ldots, p)$ and *radius* $A^r \geq 0$.

Theorem 1 *Given two spheres $\{A, B\}$ in the $\mathbb{R}^p$ space, the Hausdorff distance between A and B is given by:*

$$H(A,B) = \sqrt{\textstyle\sum_{j=1}^p \left(a_j^c - b_j^c\right)^2} + \mid A^r - B^r \mid \tag{7}$$

Proof. We remind that the equation of the sphere A is: $\sum_{j=1}^p \left(x_j - a_j^c\right)^2 = \left(A^r\right)^2$. The minimum and the maximum Euclidean distance from a point $\mathsf{O} = [x_j]$ with $(j = 1, \ldots, p)$ to the sphere A are the radii of the spheres, having centers in O, external to A and containing A, respectively. So that, the Euclidean Hausdorff distance between O and A is the minimum one.

Two spheres A and B are tangent if:

$$\textstyle\sum_{j=1}^p \left(a_j^c - b_j^c\right)^2 = (A^r \pm B^r)^2. \tag{8}$$

If A does not intersect B, we have the sign +; if A is inside B, we have the sign $-$. Let us suppose that O represents the center of the sphere B. If O belongs to A, the minimum distance between O and A is 0, obviously. If O is external to A, we need to solve the following equation for $r^\diamond$:

$$\textstyle\sum_{j=1}^p \left(a_j^c - x_j\right)^2 = (A^r + r^\diamond)^2. \tag{9}$$

Solving with respect to $r^\diamond$ we have:

$$r^\diamond = \sqrt{\textstyle\sum_{j=1}^p \left(a_j^c - x_j\right)^2} - A^r = \min_{\tilde{a} \in A} d(\mathsf{O}, A). \tag{10}$$

Let us assume that O belongs to B. The $\max\min\{d(B, A)\}$ is given by:

$$\max_{\tilde{b} \in B} \min_{\tilde{a} \in A} \left(d(\tilde{a}, \tilde{b})\right) = \max_{\tilde{b} \in B} \left(\sqrt{\textstyle\sum_{j=1}^p \left(a_j^c - \tilde{b}_j\right)^2}\right) - A^r. \tag{11}$$

Equivalently, the minimum *radius* sphere with the same center as A that both contains and is tangent to B.

According to (8) we have to solve the following equation in r^*:

$$\sum_{i=1}^p \left(a_j^c - b_j^c\right)^2 = \left(r^* - B^r\right)^2$$

$$r^* = \sqrt{\textstyle\sum_{j=1}^p \left(a_j^c - b_j^c\right)^2} + B^r. \tag{12}$$

Then the Hausdorff distance based on the Euclidean distance from a sphere B to a sphere A is:

$$h(B,A) = \max_{\tilde{b}\in B} \min_{\tilde{a}\in A} d(\tilde{b},\tilde{a}) = \sqrt{\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2} + (B^r - A^r) \qquad (13)$$

then $H(A,B) = \max(h(A,B), h(B,A)) = \sqrt{\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2} + |A^r - B^r|$ and the proof is complete. □

Given two spheres $\{A,B\}$ in the $\mathbb{R}^p$ space, $H(A,B)$ is a metric. Reflexive and symmetric properties are intuitive. We need to prove that $H(A,B) + H(B,C) \geq H(A,C)$ is true (*triangular inequality*).
For the triangular property of Euclidean (for centers) and Manhattan (for radii) distance we may assert that:

$$\sqrt{\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2} + \sqrt{\sum_{j=1}^{p} \left(b_j^c - c_j^c\right)^2} \geq \sqrt{\sum_{j=1}^{p} \left(a_j^c - c_j^c\right)^2}$$
$$|A^r - B^r| + |B^r - C^r| \geq |A^r - C^r|\,.$$

Then it follows that:

$$\sqrt{\sum_{j=1}^{p} \left(a_j^c - b_j^c\right)^2} + |A^r - B^r| + \sqrt{\sum_{j=1}^{p} \left(b_j^c - c_j^c\right)^2} + |B^r - C^r| \geq$$
$$\geq \sqrt{\sum_{j=1}^{p} \left(a_j^c - c_j^c\right)^2} + |A^r - C^r|$$

4 A comparison between two metrics

This section presents a comparison between the factorial analysis based on the two proposed measures of distance for interval valued variables. The example shows the results obtained on the "*Italian Peppers*" dataset; these data are a good example of native interval variables, they describe some chemio-physical properties (*H_2O, Glicide, Lipid, Protein*) of eight different species of Italian peppers: (*Cuban, Cuban Nano, Corno di Bue, Grosso di Nocera, Pimento, Quadrato d'Asti, Sunnybrook, Yolo wonder*). [Lauro and Palumbo, 2005]

Each factorial approach has been chosen to ensure the maximum degree of consistency with respect to the distance measure. We remind that statistical factorial analysis for interval variables does not limit itself to the study of proximities among dimensionless points but, it must take into account the size and shape of the compact subsets in $\mathbb{R}^p$

Let $[\mathbf{X}]$ be a generic $n \times p$ interval data matrix. In order to simplify the notation, we define the centers matrix $\mathbf{C} = \frac{1}{2}(\overline{\mathbf{X}} + \underline{\mathbf{X}})$ and the ranges matrix $\mathbf{R} = \frac{1}{2}(\overline{\mathbf{X}} - \underline{\mathbf{X}})$ where, $\underline{\mathbf{X}}$ and $\overline{\mathbf{X}}$ are the minimum and the maximum values matrices, respectively. All these matrices are of $n \times p$ order.

The arithmetic mean of the generic interval-valued variables $[\mathbf{x}]_j$, according to the the basic principles of the interval arithmetic [Hickey *et al.*, 2001] is defined as:

$$[\bar{\mathbf{x}}]_j = \frac{1}{n}\sum_{i=1}^{n}[x]_{i,j} = \frac{1}{n}\left[\sum_{i=1}^{n}\underline{x}_{i,j}, \sum_{i=1}^{n}\overline{x}_{i,j}\right] = \left\{\frac{1}{n}\sum_{i=1}^{n}x^c_{i,j}, \frac{1}{n}\sum_{i=1}^{n}x^r_{i,j}\right\}. \quad (14)$$

Whereas dealing with single valued variables, in $\mathbb{R}$ space, difference and distance measures are equivalent apart the sign; this is not true when variable is interval-valued. Lauro and Palumbo (2005) defined the following measure of variability for interval-valued variables based on the Hausdorff distance:

$$\text{var}([x]_j) = \frac{1}{n}\sum_{i=1}^{n}\left[\mid x^c_{i,j} - \bar{x}^c_j \mid + \mid x^r_{i,j} - \bar{x}^r_j \mid\right]^2, \quad (15)$$

where $\bar{x}^c_j$ and $\bar{x}^r_j$ represent, respectively, the mean midpoint and the mean range of the generic interval variable $[X]_j$. We call centered and reduced the interval valued variable:

$$[z]_j = \{z^c_j, z^r_j\} = \left\{\frac{(x^c_j - \bar{x}^c_j)}{\sqrt{\text{var}[x]_j}}, \frac{\mid x^r_j - \bar{x}^r_j \mid}{\sqrt{\text{var}[x]_j}}\right\}.$$

The distance presented in equation (6) can be rewritten in matrix notation as: $H^2 = \mathbf{C}\mathbf{C}^\mathsf{T} + \mathbf{R}\mathbf{R}^\mathsf{T} + \mid \mathbf{C} \mid\mid \mathbf{R}^\mathsf{T} \mid + \mid \mathbf{R} \mid\mid \mathbf{C}^\mathsf{T} \mid$, where we assume that interval variables have been centered and reduced. The quantity $trace(H^2)$ is the sum of the n squared distances from the mean. However, in the PCA practice it is preferred to apply the SVD to the $p \times p$ correlation (or covariance) matrix, in this case we will apply the SVD to the matrices $\mathbf{C}^\mathsf{T}\mathbf{C}$ and $\mathbf{R}^\mathsf{T}\mathbf{R}$. The MR-PCA of Lauro and Palumbo performs two separate PCA's on the matrices $\mathbf{C}^\mathsf{T}\mathbf{C}$ and $\mathbf{R}^\mathsf{T}\mathbf{R}$ and permits to recover the intervals on the factorial plan by adding and subtracting the rotated and translated *radii* into the space of the centers coordinates in their own space. The rotation matrix $\mathbf{T}$ is defined maximizing the quantity $\mathbf{C}^\mathsf{T}\mathbf{R}$. Notice that the square matrix $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \mathbf{C}^\mathsf{T}\mathbf{C} + \mathbf{R}^\mathsf{T}\mathbf{R} + \mid \mathbf{C}^\mathsf{T} \mid\mid \mathbf{R} \mid + \mid \mathbf{R}^\mathsf{T} \mid\mid \mathbf{C} \mid$$

is symmetric; the extra-diagonal terms vary in $[-1, 1]$ and the diagonal terms are equal to 1. It represents a sort of correlation matrix for interval valued variables, where the total correlation is the sum of three different components: midpoints association, ranges association and the midpoints/ranges congruence.

The output of the method of Palumbo and Lauro consists in a representation of the centers and of the *radii* taken into account singly, and of a joint representation where interval objects are represented by rectangles having sides parallel to the axes. However, here we propose only the midpoints and *radii* joint representation.

In order to present the second factorial approach based on the definition of the distance in (7), differently from the previous approach, we consider that both center and *radii* variables, respectively in the matrices $\mathbf{C}$ and $\mathbf{R}$, are reduced with respect to standard deviations of the centers (see [Giordani and Kiers, 2004]). Notice that the matrix B in (7) has a constant value over the main diagonal, it corresponds to the norm of the average units *radius.* The matrix notation of the distance in (7) is equal to: $\sum_{i=1}^{n} H(A_i, \bar{A})^2 = tr\,\mathbf{C}^{\mathsf{T}}\mathbf{C} + tr\,\mathbf{R}^{\mathsf{T}}\mathbf{R}$. The symbol $\bar{A}$ indicates the mean SO that is obtained by applying the formula (14). The problem consists in finding the orthogonal subspace the maximizes $tr\,\mathbf{C}^{\mathsf{T}}\mathbf{C} + tr\,\mathbf{R}^{\mathsf{T}}\mathbf{R}$ simultaneously. We introduce the super matrix $\mathbf{Y}$:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{C} \\ \mathbf{R} \end{bmatrix}.$$

The projection of $\mathbf{Y}$ on a common orthogonal subspace can be obtained by means of the extraction of the principal components of $\mathbf{C}^{\mathsf{T}}\mathbf{C}$ denoted as $\mathbf{D_{CC}}$. Considering the projection of $\mathbf{Y}$ on the space spanned by the centers using the projector $\mathbf{P_C}$ and on the orthogonal projection using $\mathbf{P}_{\mathbf{C}}^{\perp}$ we have:

$$\begin{aligned} \mathbf{Y} &= \mathbf{P_C Y} + \mathbf{P}_{\mathbf{C}}^{\perp}\mathbf{Y} = (\mathbf{P_C C}, \mathbf{P_C R}) + (\mathbf{P}_{\mathbf{C}}^{\perp}\mathbf{C}, \mathbf{P}_{\mathbf{C}}^{\perp}\mathbf{R}) = \\ &= (\mathbf{C}, \mathbf{P}_{\mathbf{C}}^{\perp}\mathbf{R}) + (\mathbf{0}, \mathbf{P}_{\mathbf{C}}^{\perp}\mathbf{R}) \end{aligned} \tag{16}$$

that leads to the following decomposition:

$$\mathbf{D_{YY}} = \begin{bmatrix} \mathbf{D_{CC}} & \mathbf{D_{CR}} \\ \mathbf{D_{RC}} & \mathbf{D_{RR}} \end{bmatrix} = \begin{bmatrix} \mathbf{D_{CC}} & \mathbf{D_{CR}} \\ \mathbf{D_{RC}} & \mathbf{D_{RRC}} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{D}_{\mathbf{RR}.\frac{\perp}{\mathbf{C}}} \end{bmatrix} \tag{17}$$

where $\mathbf{D}_{\mathbf{RR}.\frac{\perp}{\mathbf{C}}}$ can be obtained computing the first principal components of $\mathbf{C}$ and then obtaining the structure matrix of $\mathbf{D_{CC}}$ on the base of the set of the principal components of $\mathbf{D_{CC}}$. For further details see [Takeuchi *et al.*, 1982]. Taking into account these results, there are several possible approaches to the analysis [Lebart *et al.*, 1995]; for sake of space, here we do not discuss the choices and their motivations. Figure 1 shows the results based on the distance between *boxes*: ranges are rotated and projected into the space of the midpoints as supplementary points. The total variability associated to the first factorial plan is 65.76%. Figure 2 shows the results obtained representing the SO with respect to the distance between hyperspheres. Here the variability associated to the first factorial plan is equal to 76.48%.

Looking at the outputs we notice that the SO can be distinguished according to their position, size and shape. It is worth noticing that, with respect to the positions, results of the two analyses are consistent. SO have the same order on the first factor in both analyses. The size interpretation is quite intuitive. Few words are necessary to correctly interpret the shapes: it is necessary to take into account the shape itself and also the range orientation. Looking at figure 1, we notice that *Sunnybrook* and *Quadrato d'Asti* have similar shapes but they have opposite range orientation. This is confirmed

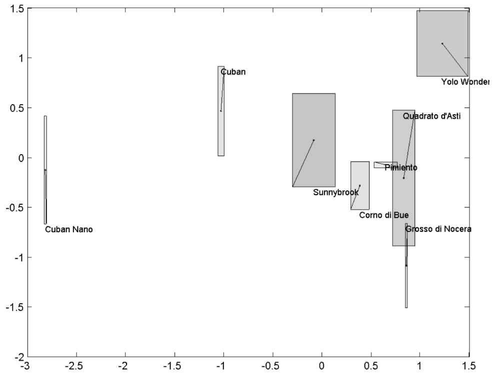

Fig. 1. Distance between *boxes*: first factorial plan (65.76%)

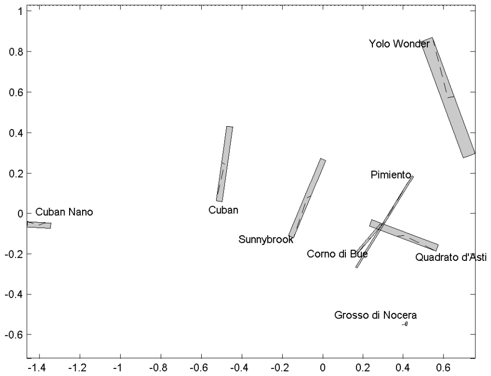

Fig. 2. Distance between hyperspheres: first factorial plane (76.48%)

also in the other analysis: in figure 2 we see that *Sunnybrook* and *Quadrato d'Asti* appear orthogonal.

To understand which variables have mainly characterized the positioning and the size and shape of the SO it is necessary to look at the variables representations on the same factorial plans.

5 Conclusion and future work

Since Edwin Diday introduced Symbolic Data Analysis [Diday, 1989] we have noticed a growing interest for the analysis of complex data structures. The first book entirely dedicated to Symbolic Data Analysis appeared five years ago [Bock and Diday, 2000]. These new statistical data need new concepts not having a counterpart in the "classical" data analysis, necessarily. At the beginning, many have proposed special data-codings to make data tractable by the traditional methods; so that, most of the big effort done up to now allowed us to treat complex data with *suitably adapted* methods for single-valued data. We believe that the next challenge is to setup numerical and statistical methods that are specifically designed for the complex-data structures. We see two main research directions: *i*) definition of new statistical indexes (measures of central tendency, variability, etc.) that take into account the innovative nature of the data; *ii*) development of analytical and numerical methods allowing to treat intervals as mathematical structures. Interval arithmetic has been mainly developed to treat data imprecision caused by the "old" fix-point CPU (the round-off error) and its generalization to the statistical interval-valued data requires a big effort.

Nevertheless, the treatment of set-valued variables is a field with very high potential for further developments.

References

[Alefeld and Mayer, 2000]G. Alefeld and G. Mayer. Interval analysis: theory and applications. *Jour. of Comp. and Appl. Mathematics*, 121:421–464, 2000.

[Bock and Diday, 2000]H. H. Bock and E. Diday, eds. *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data.* Springer Verlag, Heidelberg, 2000.

[Braun *et al.*, 2003]D. Braun, J. Mayberry, A. Powers, and S. Schlicker. The geometry of the Hausdorff metric. Available at: `http://faculty.gvsu.edu/schlicks/Hausdorff_ paper.pdf`, August 2003.

[Cazes *et al.*, 1997]P. Cazes, A. Chouakria, E. Diday, and Y. Schektman. Extension de l'analyse en composantes principales à des données de type intervalle. *Revue de Statistique Appliquée*, XIV(3):5–24, 1997.

[Diday, 1989]E. Diday. Introduction à l'approche symbolique en analyse des données. *Rev. d'Aut., d'Informatique et de Rec. Opérationnelle*, 32(2), 1989.

[Giordani and Kiers, 2004]P. Giordani and H.A.L. Kiers. Principal component analysis of symmetric fuzzy data. *Comp. Stat. Data Anal.*, 45:519–548, 2004.

[Hickey *et al.*, 2001]T. Hickey, Q. Ju, and M. H. Van Emden. Interval arithmetic: From principles to implementation. *Jour. of the ACM*, 48(5):1038–1068, 2001.

[Lauro and Palumbo, 2000]C. N. Lauro and F. Palumbo. Principal component analysis of interval data: A symbolic data analysis approach. *Computational Statistics*, 15(1):73–87, 2000.

[Lauro and Palumbo, 2005]C. N. Lauro and F. Palumbo. Principal component analysis for non-precise data. In M. Vichi *et al.*, eds, *New Developments in Classification and Data Analysis*, pages 173–184. Springer, 2005.

[Lebart *et al.*, 1995]Ludovic Lebart, Alain Morineau, and M. Piron. *Statistique exploratorie multidimensionelle.* Dunod, Paris, 1995.

[Neumaier, 1990]A. Neumaier. *Interval methods for systems of Equations.* Cambridge University Press, Cambridge, 1990.

[Takeuchi *et al.*, 1982]K. Takeuchi, H. Yanai, and B. N. Mukherjee. *The Fundations of Multivariate Analysis.* Wiley Eastern Ltd., New Delhi, 1982.

[Xiang *et al.*, 2004]G. Xiang, S. A. Starks, V. Kreinovich, and L. Longpre. New algorithms for statistical analysis of interval data. Utep-cs-04-04, NASA PACES, El Paso, TX 79968, USA, 2004. at: `http://www.cs.utep.edu/vladik/2004/list04.html`.